University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2009 *Paper* 81

WEIGHTING AND PREDICTION IN SAMPLE SURVEYS

Rod Little*

*rlittle@umich.edu

WEIGHTING AND PREDICTION IN SAMPLE SURVEYS

Rod Little

Abstract

A fundamental technique in survey sampling is to weight included units by the inverse of their probability of inclusion, which may be known (as in the case of sampling weights) or estimated (as in the case of nonresponse weights). The technique is closely associated with the design-based approach to survey inference, with the idea that units in the sample are representing a certain number of units in the population. I discuss weighting from a modeling perspective. Some common misconceptions of weighting will be addressed, including the idea that modelers can ignore the sampling weights, or that weighting necessarily reduces bias at the expense of increased variance, or that units entering the calculation of nonresponse weights should be weighted by their sampling weights. A robust model-based perspective suggests that selection weights cannot be ignored, but there may be better ways of incorporating them in the inference than via the standard Horvitz-Thompson estimator and its variants.

Revised Draft

February 23, 2009

Weighting and prediction in sample surveys

Roderick J. Little

University of Michigan

Abstract

A fundamental technique in survey sampling is to weight included units by the inverse of their probability of inclusion, which may be known (as in the case of sampling weights) or estimated (as in the case of nonresponse weights). The technique is closely associated with the design-based approach to survey inference, with the idea that units in the sample are representing a certain number of units in the population. I discuss weighting from a modeling perspective. Some common misconceptions of weighting will be addressed, including the idea that modelers can ignore the sampling weights, or that weighting necessarily reduces bias at the expense of increased variance, or that units entering the calculation of nonresponse weights should be weighted by their sampling weights. A robust model-based perspective suggests that selection weights cannot be ignored, but there may be better ways of incorporating them in the inference than via the standard Horvitz-Thompson estimator and its variants.

Keywords: Bayesian methods; design-based inference; sampling weights; regression; robustness; survey sampling

1. Introduction

It is an honor to write an article in celebration of the diamond jubilee of the Calcutta Statistical Association Bulletin, a venerable statistical institution, and to acknowledge the profound contribution of Indian statisticians to progress in our field. Historically, this is clear when we consider the influence of major Indian statisticians like Basu, Gnanadesikan, Mahalanobis, and more recently C.R. Rao, not to mention the distinguished Rao's with other initials, and many others. Personally, my career has been

enhanced by numerous friendships and encounters with Indian statisticians; my boss in my first real job at World Fertility Survey was the demographer VC Chidambaram (Chid to all who knew him) who was a sympathetic colleague and strong leader; another fine colleague at World Fertility Survey was Vijay Verma, an outstanding student of Leslie Kish who played a leading role in sampling activities in that large study. More recently, I have since collaborated extensively with my colleague Trivellore Raghunathan at Michigan, on topics of sampling inference and missing data. Indeed Biostatistics at Michigan has a strong Indian representation in terms of faculty and students.

I write about the role of weights in the analysis of survey samples. Probability sampling is one of the key contributions of statistics, and this is an area where Indian statisticians have made seminal contributions (e.g. Mahalanobis 1943; Godambe 1955; Basu 1971; Rao 1997, 2003). Many of the key aspects of probability sampling, including stratification and multistage sampling, were first implemented on a large scale in India. It has interested me since my time working at the World Fertility Survey, where the virtues of probability sampling were widely touted by Sir Maurice Kendall and Leslie Kish, and the question of making analytic inferences that incorporated the survey design was of great interest. As a statistician drawn to the Bayesian paradigm for survey inference, sample surveys are a challenge since the prevailing paradigm of survey sample inference is design-based, and survey samplers have a widespread distrust of models.

2. Survey weighting, prediction, and design vs. model-based inference.

The clash between two approaches to weighting survey data puzzled me as a student of statistics. Early on we learn about linear regression, fitted by ordinary least squares (OLS), which is optimal for a model that assumes that the residual variance is

constant for all values of the covariates. If the variance of the residual for unit *i* is σ^2 / u_i for some known constant u_i, then better inferences are obtained by weighted least squares, with unit *i* assigned a weight proportional to u_i. This form of weighting is model-based, since the linear regression model for the outcome (say *Y*) has been modified to incorporate a non-constant residual variance.

Later I took a course in survey sampling, and learnt about a different form of weighting, based on the selection probabilities. If unit *i* is sampled with selection probability π_i, then the survey sampler replaces OLS by weighted least squares, weighting the contribution of unit *i* to the least squares equations by $w_i \propto 1/\pi_i$, the inverse of the probability of selection. This form of weighting is design-based, with π_i relating to the selection of units: since unit *i* "represents" $1/\pi_i$ units of the population, it receives a weight proportional to $1/\pi_i$ in the regression.

Both forms of weighting seem plausible, but they are not necessarily the same. So which is correct? The answer is not obvious -- the role of sampling weights in regression has been extensively debated in the literature –see for example Konijn (1962), Brewer and Mellor (1973), Dumouchel and Duncan (1983), Smith (1988), Little (1991), Pfeffermann (1993), Korn and Graubard (1999). In fact, it rests fundamentally on whether one adopts a design-based on model-based perspective on statistical inference.

The design-based approach to survey inference (e.g. Hansen, Hurwitz and Madow 1953, Kish 1965, Cochran 1977) has the following main features. For a population with *N* units, let $Y = (y_1, ..., y_N)$ where y_i is the set of survey variables for unit *i*, and let $I = (I_1, ..., I_N)$ denote the set of *inclusion indicator variables,* where $I_i = 1$ if unit *i* is

included in the sample and $I_i = 0$ if it is not included. Design-based inference for a finite population quantity $Q = Q(Y)$ involves the choice of an estimator $\hat{q} = \hat{q}(Y_{\text{inc}}, I)$, a function of the observed part Y_{inc} of *Y*, that is unbiased or approximately unbiased for *Q* with respect to the distribution *I*; and the choice of a variance estimator $\hat{v} = \hat{v}(Y_{\text{inc}}, I)$ that is unbiased or approximately unbiased for the variance of $\hat{q}$ with respect to the distribution of *I*. Inferences are then generally based on normal large sample approximations. For example, a 95% confidence interval for *Q* is $\hat{q} \pm 1.96\sqrt{\hat{v}}$.

The model-based approach to inference bases inference on the distribution of *Y*, and usually does not overtly consider a distribution for *I*; while assumptions of randomization lurk in the background, they are not the basis for the inference. The model for the survey outcomes *Y* is used to predict the non-sampled values of the population, and hence finite population quantities *Q*. There are two major variants: superpopulation modeling and Bayesian modeling. In superpopulation modeling (e.g. Royall 1970; Thompson 1988; Valliant, Dorfman, and Royall 2000), the population values of *Y* are assumed to be a random sample from a "superpopulation", and assigned a probability distribution $p(Y \mid \theta)$ indexed by fixed parameters θ. Bayesian survey inference (Ericson 1969, 1988; Basu 1971; Scott 1977; Binder 1982; Rubin 1983, 1987; Ghosh and Meeden 1997, Little 2004) requires the specification of a prior distribution $p(Y)$ for the population values. Inferences for finite population quantities $Q(Y)$ are then based on the posterior predictive distribution $p(Y_{\text{exc}} \mid Y_{\text{inc}})$ of the non-sampled values (say Y_{exc}) of *Y*, given the sampled values Y_{inc}. The specification of the prior distribution $p(Y)$ is often

achieved via a parametric model $p(Y \mid \theta)$ indexed by parameters θ, combined with a prior distribution $p(\theta)$ for θ, that is:

$$p(Y) = \int p(Y \mid \theta) p(\theta) d\theta .$$

The posterior predictive distribution of Y_{exc} is then

$$p(Y_{\text{exc}} \mid Y_{\text{inc}}) \propto \int p(Y_{\text{exc}} \mid Y_{\text{inc}}, \theta) p(\theta \mid Y_{\text{inc}}) d\theta$$

where $p(\theta \mid Y_{\text{inc}})$ is the posterior distribution of the parameters, computed via Bayes' Theorem:

$$p(\theta \mid Y_{\text{inc}}) = p(\theta) p(Y_{\text{inc}} \mid \theta) / p(Y_{\text{inc}}),$$

where $p(\theta)$ is the prior distribution, $p(Y_{\text{inc}} \mid \theta)$ is the likelihood function, viewed as a function of θ, and $p(Y_{\text{inc}})$ is a normalizing constant. This posterior distribution induces a posterior distribution $p(Q \mid Y_{\text{inc}})$ for finite population quantities $Q(Y)$.

The specification of $p(Y \mid \theta)$ in this Bayesian formulation is the same as in parametric superpopulation modeling, and in large samples the likelihood based on this distribution dominates the contribution from the prior for θ. As a result, large-sample inferences from the superpopulation modeling and Bayesian approaches are often similar.

Example 1. Estimating a mean from a stratified sample.

Consider the simple case of estimation of a finite population mean $\bar{Y}$ from a stratified random sample. Suppose the population is divided into *J* strata, and let N_j be the known population count in stratum *j* and $\bar{Y}_j$ the unknown population mean in stratum *j*. The quantity of interest is $Q = \bar{Y} = \sum_{j=1}^{J} P_j \bar{Y}_j$, where $P_j = N_j / N$ is the proportion of the

population in stratum *j*. We assume that a random sample of size n_j of the N_j units are sampled in stratum *j*, and let $\{y_{ji}, i = 1,...,n_j\}$ denote the set of sampled *Y*-values in stratum *j*. Then $Y_{\text{inc}} = \{y_{ji}, j = 1,...,J; i = 1,...,n_j\}$. Stratified random sampling is defined by:

$$\Pr(I_{ji} = 1) = \left[\binom{N_j}{n_j}\right]^{-1}, \text{ if } \sum_{i=1}^{N_j} I_{ji} = n_j, \text{ and 0 otherwise.}$$

The usual estimator of $\overline{Y}$ in this setting is the stratified mean

$$\hat{q} = \overline{y}_{\text{st}} \equiv \sum_{j=1}^{J} P_j \overline{y}_j = \left(\sum_{j=1}^{J} n_j \overline{y}_j / \pi_j\right) / \left(\sum_{j=1}^{J} n_j / \pi_j\right), \tag{1}$$

where $\overline{y}_j$ is the sample mean in stratum *j*. The estimator (1) is the weighted mean of the sampled units, where units in stratum *j* are weighted by the inverse of their selection probability $\pi_j = n_j / N_j$.

Consider now a model-based approach. Suppose we assume the model

$$y_{ji} \sim_{\text{ind}} \text{Nor}(\mu, \sigma^2 / u_j) \tag{2}$$

where Nor(*a*,*b*) denotes the normal distribution with mean *a*, variance *b*, u_j is known, and the non-informative prior distribution

$$p(\mu, \log \sigma^2) = \text{const.} \tag{3}$$

The posterior mean of the population total is

$$\overline{y}_{\text{u}} = \left(\sum_{j=1}^{J} n_j u_j \overline{y}_j\right) / \left(\sum_{j=1}^{J} n_j u_j\right), \tag{4}$$

which weights cases in stratum *j* by u_j, rather than $1/\pi_j$.

The application of design weights in this example is not controversial, and the stratified mean is difficult to beat as an estimator except in unusual situations. Indeed, the model-based estimator (4) is not recommended, since it is vulnerable to the assumption that the stratum means are equal. If the model (2)-(3) is changed to allow a separate mean in each stratum:

$$y_{ji} \sim_{\text{ind}} \text{Nor}(\mu_j, \sigma^2 / u_j) \tag{5}$$

$$p(\mu_j, \log \sigma^2) = \text{const.}, \tag{6}$$

the posterior mean is then the stratified mean (1), so the design and model-based estimates correspond. Usually allowing a separate mean in each stratum is sensible, since strata are generally chosen to be related to survey outcomes; we do not determine strata by the toss of a coin.

In other settings, the design-weighted Horvitz-Thompson estimator (Horvitz and Thompson 1952) can lead to nonsensical estimates. Basu (1971) gave the following famous and amusing example:

Example 2. Basu's elephants. "The circus owner is planning to ship his 50 adult elephants and so he needs a rough estimate of the total weight of the elephants. As weighing an elephant is a cumbersome process, the owner wants to estimate the total weight by weighing just one elephant. Which elephant should he weigh? So the owner looks back on his records and discovers a list of the elephants' weights taken 3 years ago. He finds that 3 years ago Sambo the middle-sized elephant was the average (in weight) elephant in his herd. He checks with the elephant trainer who reassures him (the owner) that Sambo may still be considered to be the average elephant in the herd. Therefore, the owner plans to weigh Sambo and take 50*y* (where *y* is the present weight of Sambo) as an

estimate of the total weight $Y = Y_1 + Y_2 + ...Y_{50}$ of the 50 elephants. But the circus statistician is horrified when he learns of the owner's purposive sampling plan. "How can you get an unbiased estimate of *Y* this way?" protests the statistician. So, together they work out a compromise sampling plan. With the help of a table of random numbers they devise a plan that allots a selection probability of 99/100 to Sambo and equal selection probabilities of 1/4900 to each of the other 49 elephants. Naturally, Sambo is selected and the owner is happy. "How are you going to estimate *Y?"*, asks the statistician. "Why? The estimate ought to be 50*y* of course," says the owner. "Oh! No! That cannot possibly be right," says the statistician, "I recently read an article in the *Annals of Mathematical Statistics* where it is proved that the Horvitz-Thompson estimator is the unique hyperadmissible estimator in the class of all generalized polynomial unbiased estimators."What is the Horvitz-Thompson estimate in this case?" asks the owner, duly impressed. "Since the selection probability for Sambo in our plan was 99/100," says the statistician, "the proper estimate of *Y* is 100*y*/99 and not 50*y*." "And, how would you have estimated *Y*," inquires the incredulous owner, "if our sampling plan made us select, say, the big elephant Jumbo?" "According to what I understand of the Horvitz-Thompson estimation method," says the unhappy statistician, "the proper estimate of *Y* would then have been 4900*y,* where *y* is Jumbo's weight." That is how the statistician lost his circus job (and perhaps became a teacher of statistics!)"

Design-based statisticians groan when modelers bring up Basu's example, since they view it as a caricature: no sensible design-based statistician would use the HT estimator in this case. Basu was using the example to make a theoretical point; the HT estimator has the useful property of design-unbiasedness in large samples, but no single

estimator is optimal in all situations, and weighted estimators can do very badly, particularly in small samples. As a more realistic example, design-based statisticians deviate from strict weighting when outlying observations receive large weights, and dominate the estimator.

Slavish adoption of the design-weighted estimator without attention to whether the underlying model is reasonable is not wise. How can we tell when the HT estimator is not going to work? One approach is to consider the model for the population implied model by weighting. Specifically, consider creating an estimate of the population by replicating sample observation i $1/\pi_i$ times. Is the resulting population sensible as an approximation for the problem at hand? Clearly the answer is "yes" in Example 1, and "no" in Example 2. When the answer is no, better estimates exist.

The population that replicates the sample is a kind of model, and design-based statisticians cannot avoid models. On the other hand, model-based statisticians cannot avoid weights, since a model that ignores the survey weights is likely to be poorly calibrated, given the realities of model misspecification as exemplified by the absence of stratum means in (2). For other examples, see Kish & Frankel (1974), Hansen, Madow & Tepping (1983), Holt, Smith, and Winter (1980), and Pfeffermann and Holmes (1985).

My own philosophy of survey sampling inference, as for statistics in general, is calibrated Bayes, where inferences are Bayesian and based on models for *Y*, but models need to be calibrated in the sense of having good design-based properties in repeated sampling from the distribution of *I* (Box 1980, Rubin 1984, Little 2006). The calibrated Bayes philosophy leads to prediction models with relatively noninformative prior distributions, which incorporate design features appropriately, seeking both efficiency

and robustness to model misspecification. My work in this area has been guided by this underlying principle.

For calibrated Bayesians, both the distribution of *Y* and the distribution of *I* are important – indeed a useful and unifying conceptual device is to formulate the model in terms of the joint distribution of both *Y* and *I*. The early literature of surveys focused either on the distribution of *Y* or the distribution of *I*, rather than the joint distribution of *Y* and *I*. This tended to lead to compartmentalization into design-based and model-based advocates. To my knowledge, the first person to explicitly model *I* and *Y* seems to be Rubin (1978), in a paper that was more focused on estimating treatment effects but also modeled the selection mechanism.

The joint modeling of *Y* and *I* in the survey context is well described in the book by Gelman et al. (2003). The following description is from Little (2003a). The model can be formulated as:

$$p(y_U, i_U \mid z_U, \theta, \phi) = p(y_U \mid z_U, \theta) \times p(i_U \mid z_U, y_U, \phi),$$

where *U* denotes universe as opposed to sample, y_U denotes the survey data, i_U the sample inclusion indicators, z_U denotes design variables, such as strata indicators, and θ, ϕ are unknown parameters. The likelihood of θ, ϕ based on the observed data (z_U, y_{inc}, i_U) is then:

$$L(\theta, \phi \mid z_U, y_{inc}, i_U) \propto p(y_{inc}, i_U \mid z_U, \theta, \phi) = \int p(y_U, i_U \mid z_U, \theta, \phi) dy_{exc}.$$

The more usual likelihood does not include the inclusion indicators i_U as part of the model. Specifically, the likelihood *ignoring the selection process* is based on the model for y_U alone:

$$L(\theta \mid z_U, y_{inc}) \propto p(y_{inc} \mid z_U, \theta) = \int p(y_U \mid z_U, \theta) dy_{exc} .$$

Applying Rubin's (1976) theory, sufficient conditions for ignoring the selection mechanism are:

Selection at Random (SAR): $p(i_U \mid z_U, y_U, \phi) = p(i_U \mid z_U, y_{inc}, \phi)$ for all y_{exc}.

Distinctness: θ, ϕ have distinct parameter spaces

Probability sample designs are generally both ignorable and known, in the sense that:

$$p(i_U \mid z_U, y_U, \phi) = p(i_U \mid z_U, y_{inc}),$$

where z_U represents known sample design information, such as clustering or stratification information. Thus the sampling mechanism can be ignored, provided the sample design information in z_U is included in the model. In the case of weighting, this means conditioning on the design variables that lead to differential weights. This analysis also provides a justification for randomization in design, since other forms of sampling, like quota sampling or purposive selection, do not necessarily satisfy the SAR assumption. Extensions to handle survey nonresponse are given in Little (1982, 2003b).

The sampling weights in Examples 1 and 2 are determined solely by the probabilities of selection. More generally, survey weights also involve components for survey nonresponse and for post-stratification to match known population distributions. The standard approach creates a composite weight for unit *i* of the form

$$w_i \propto w_{is} \times w_{in}(w_{is}) \times w_{ip}(w_{is}, w_{in}) \qquad (7)$$

where w_{is} is the sampling weight, $w_{in}(w_{is})$ is a nonresponse weighting factor and $w_{ip}(w_{is}, w_{in})$ is a post-stratification adjustment. In the remainder of this article I'll give

some additional illustrations of prediction models for samples with features like selection probabilities and survey nonresponse.

3. Weights that incorporate population information

In Example 1 we noted that the weighting and prediction approaches yield the stratified mean in the case of stratified example. Post-stratification is a closely related example:

Example 3. Inference for the mean with categorical post-strata. Another situation where the design and model-based approaches intersect is estimation of the population mean of a variable *Y* from a simple random sample, given a categorical post-stratum variable *Z* with known distribution in the population. Let y_{ji} denote the value of *Y* for sampled unit *i* in post-stratum $Z = j$. Assume the model of Eqs. (5) - (6). The posterior distribution of the population mean has mean

$$\overline{y}_{\text{mod}} = \overline{y}_{\text{wt}} = \sum_{j=1}^{J} P_j \overline{y}_j = \sum_{j=1}^{J} w_j n_j \overline{y}_j \,/\, \sum_{j=1}^{J} w_j n_j \,, \tag{8}$$

where in post-stratum $Z = j$, P_j is the population proportion, n_j is the sample size, $\overline{y}_j$ is the sample mean, and $w_j = nP_j / n_j$. The estimate (8) is the post-stratified mean, also obtained in the design-based approach by applying post-stratification weights w_j to the sampled units in post-stratum *j*.

Asymptotically (8) works fine, but in small samples it is unstable. The situation here differs from stratification on *Z*, where the stratum counts $\{ n_j \}$ are under the control of the sampler. With post-stratification, the $\{ n_j \}$ are determined by which units happen

to fall into post-stratum *j*. The post-stratum counts n_j in one or more post-strata may become very small, yielding large weights w_j; indeed (4) is not defined if for any *j* $n_j = 0$, and it does not have a well-defined sampling distribution in repeated samples unless $\{n_j\}$ are constrained to be positive; for discussion of this point see Holt and Smith (1979) and Little (1993). Design-based approaches modify the weights, for example by pooling small post-strata. However, from a prediction perspective, the problem lies not in the weights, but in the unstable predictions $\bar{y}_j$ of the means in post-strata with small counts. The associated proportions P_j are, after all, known!

From a Bayesian perspective, the posterior distribution of $\bar{Y}$ for the model (5) – (6). is a mixture of t distributions, and as such incorporates t corrections from estimating the variance that are not available under the design-based approach, which is basically asymptotic. Concerning the instability of (8), the Bayesian solution is to modify the prior distribution (6) to allow borrowing of strength across post-strata. One such modification is

$$\mu_j \sim_{\text{ind}} N(\mu, \tau^2), p(\mu, \log\sigma^2, \tau^2) = \text{const.},$$

which yields predictions that effectively shrink the weights w_j to a constant. This approach to weight shrinkage is discussed in Little (1993), and extensions in the presence of covariates are discussed in Lazzeroni and Little (1998) and Elliott and Little (2000).

Example 4. Categorical strata and post-strata. Suppose now that we have a stratified sample, with stratifier Z_1 with population distribution $\{P_{1j}, j = 1,..., J\}$, and we also know the population distribution $\{P_{2k}, k = 1,..., K\}$ of a post-stratification variable Z_2. The traditional weighting approach (7) is to post-stratify the stratification weights so that the

weighted sample counts match the population distribution of Z_2. That is, the composite weight for units in stratum *j*, post-stratum *k* is

$$w_{jk} = w_{1j} \times w_{2k \cdot j},$$

where $w_{1j} = nP_{1j} / n_{1j}$ and $w_{k \cdot j} = nP_{2k} w_{1j} / \sum_{\ell} w_{1\ell}$. Interestingly, these weights lead to stratum counts that do not match the population distribution of Z_1. From a modeling perspective, the data about the joint distribution of Z_1 and Z_2 consists of the sample counts $\{n_{jk}\}$ and the known marginal distributions of Z_1 and Z_2. A saturated model for the joint distribution of *Y*, Z_1 and Z_2 takes the form:

$$\begin{aligned} &\{n_{jk}\} \sim \text{MNOM}(n, P_{jk}); \\ &y_{jki} \sim \text{Nor}(\mu_{jk}, \sigma^2_{jk}),\ p(\mu_{jk}, \log \sigma^2_{jk}) = \text{const.} \end{aligned} \tag{9}$$

Maximum likelihood estimates $\{\hat{P}_{jk}\}$ of $\{P_{jk}\}$ are obtained by raking the sample counts to match the Z_1 and Z_2 margins by iterative proportional fitting, yielding weights that match both of these margins. The maximum likelihood estimate of the population mean of *Y* is then

$$\bar{y}_{\text{mod}} = \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{P}_{jk} \bar{y}_{jk}\ . \tag{10}$$

Classification by both Z_1 and Z_2 increases the likelihood of small counts $\{n_{jk}\}$ in some cells, so modifications of (9) for predicting the cell means may be important. One possibility is to replace the saturated model by

$$\begin{aligned} &y_{jki} \sim \text{Nor}(\mu + \alpha_j + \beta_k + \gamma_{jk}, \sigma^2_{jk}), \\ &\sum_{j=1}^{J} \alpha_j = \sum_{k=1}^{K} \beta_k = 0, \gamma_{jk} \sim \text{Nor}(0, \tau^2) \end{aligned} \tag{11}$$

which results in shrinkage of the sample mean $\bar{y}_{jk}$ towards the fitted mean for the additive model relating Y to Z_1 and Z_2. In summary, adopting a prediction perspective (a) corrects the usual estimator to match both stratum and post-stratum margins; (b) provides t corrections for estimating the variance, as in Example 3; and (c) allows modifications of the estimator (10) in small samples by modifying the prior distribution of the cell means.

Example 5. Probability proportional to size (PPS) sampling.

The weights in Examples 3 and 4 incorporate information from categorical variables in the population. Sometimes sample designs involve stratifiers that are continuous variables. A common design with a continuous stratifier is PPS sampling, where units are selected with probability proportional to a size variable *Z* known for all units in the population. The standard design-based estimator in this setting is the HT estimator

$$\bar{y}_{\text{wt}} = \frac{1}{N}\left(\sum_{i=1}^{n} y_i / \pi_i\right) \tag{12}$$

where π_i is the probability of selection for unit *i*. From a modeling perspective, the objective is to base estimates on predictions from a regression model for the distribution of *Y* given *Z*. The estimator (12) is approximately the prediction estimator for the "HT model"

$$y_i \mid z_i \sim \text{Nor}(\beta z_i, \sigma^2 z_i^2)\,. \tag{13}$$

The estimator (12) tends to be efficient when the HT is satisfied, but does poorly when this model is seriously violated. Zheng and Little (2003, 2004, 2005) consider predicting the non-sampled values using the a more flexible penalized spline model

$$y_i \sim \text{Nor}(f(z_i, \beta), \sigma^2 z_i^k),$$

where f is a spline function:

$$f(z_i, \beta) = \beta_0 + \sum_{j=1}^{p} \beta_j z_i^j + \sum_{l=1}^{m} \beta_{l+p} (z_i - \kappa_l)_+^p, i = 1, ..., N.$$

Here $k \geq 0$ is a constant reflecting the knowledge of the error variance and the constants $\kappa_1 < ... < \kappa_m$ are selected fixed knots, and $(u)_+^p = u^p$ if $u > 0$ and 0, otherwise; and $(\beta_{p+1}, ..., \beta_{p+m})^T$ are assumed $\text{Nor}(0, \tau^2 I_m)$. This model relaxes the assumption that the relationship between Y and Z is linear. Zheng and Little (2005) show by simulation that prediction inferences based on this model yield gains over the HT estimator in both efficiency and confidence coverage when the HT model (13) is violated, while sacrificing little in terms of efficiency when the HT model is satisfied. Chen, Elliott and Little (2008) develop Bayesian inference for a population proportion from unequal probability samples, where the probit of the probability that $y_i = 1$ is modeled as penalized spline of the size variable. They also show gains in terms of efficiency and confidence coverage compared with the HT estimator, and generalized regression extensions of the HT estimator.

4. Unit and Item Nonresponse

In the context of survey nonresponse, weighting adjustments are common in the case of unit nonresponse, as in the following example.

Example 6. Unit nonresponse in surveys

Suppose that respondents and nonrespondents are classified into *C* adjustment cells based on covariates *X* observed for both. The nonresponse weight in cell *c* is then

the inverse of the estimated response rate in that cell. This is also the prediction estimator for a model that assumes a different mean for the outcome in each adjustment cell. Some comments on this approach follow:

(1) Given extensive covariate information, adjustment cells should be chosen that are predictive of both the survey outcomes and of nonresponse. Adjustment cell weighting, and extensions based on models for the propensity to respond, tend to focus on good predictors of response, but Little and Vartivarian (2005) argue that having a good predictor of the outcome is more important; these can actual improve efficiency of estimation, and good predictors of nonresponse that are not related to the outcome simply increase variance without reducing bias.

2. When the sampling weights are not constant within adjustment cells, it is common practice to compute the nonresponse weight as the inverse of the weighted response rate, where units are included in the rate weighted by their sampling weights. This "weight squared" approach does not correct for bias when the outcome is related both to the adjustment cell variable and the stratification variable, as is demonstrated by simulations in Little and Vartivarian (2003).

3. Since nonresponse is not under the control of the sampler, highly variable nonresponse weights are possible, as when the fraction of respondents in an adjustment cell is small. Thus shrinkage of the nonresponse weights may be attractive, and this is accomplished by putting a proper prior on the adjustment cell means, as was done in Example 3 in the case of post-stratification.

Example 7. Item nonresponse in surveys.

Item nonresponse occurs when particular items in the survey are missing, because they were missed by the interview, or the respondent declined to answer particular questions. For item nonresponse the pattern of missing values is general complex and multivariate, and substantial covariate information is available to predict the missing values in the form of observed items. These characteristics make weighting adjustments unattractive, since weighting methods are difficult to generalize to general patterns of missing data (Little 1988) and make limited use of information in the incomplete cases.

A common practical approach to item missing data is imputation, where missing values are filled in by estimates and the resulting data are analyzed by complete-data methods. In this approach incomplete cases are retained in the analysis. Imputation methods until the late 1970's lacked an underlying theoretical rationale. Pragmatic estimates of the missing values were substituted, such as unconditional or conditional means, and inferences based on the filled-in data. A serious defect with the method is that it "invents data". More specifically, a single imputed value cannot represent all of the uncertainty about which value to impute, so analyses that treat imputed values just like observed values generally underestimate uncertainty, even if nonresponse is modeled correctly. Rubin's (1987) theory of multiple imputation (MI) put imputation on a firm theoretical footing, and also provided simple ways of incorporating imputation uncertainty into the inference. Instead of imputing a single set of draws for the missing values, a set of Q (say $Q = 10$) datasets are created, each containing different sets of draws of the missing values from their predictive distribution given the observed data. The analysis of interest is then applied to each of the Q datasets and results are combined

using simple multiple imputation combining rules (Rubin 1987; Little and Rubin, 2002). An alternative to multiple imputation is to use sample re-use methods that reimpute the data on each replicate sample (Rao 1996).

5. Conclusion

The above examples suggest that weighting provides a useful all-purpose approach to large sample estimation in surveys, but Bayesian predictive models yield useful extensions and refinements, provided careful attention is paid to incorporating the survey design. Some advantages of the Bayesian approach are:

(1) it provides a unified approach to survey inference, aligned with mainline statistics approaches in other application areas such as econometrics.

(2) In large samples and with uninformative prior distributions, results can parallel those from design-based inference, as we have seen in the case of stratified and post-stratified sampling in Examples 1 and 3.

(3) The Bayesian approach is well equipped to handle complex design features such as clustering through random cluster models (Scott and Smith 1969), stratification through covariates that distinguish strata, nonresponse (Little 1982; Rubin 1987; Little and Rubin 2002) and response errors.

(4) The Bayesian approach may yield better inferences for small sample problems where exact frequentist solutions are not available, by propagating error in estimating parameters. For example, the posterior distribution of the mean for inference from normal stratified samples in Example 3 is a mixture of t distributions that propagates uncertainty in estimating the stratum variances. On

the other hand, the standard design-based inference based on the normal distribution assumes that the stratum variances are estimated without error from the sample.

(5) The Bayesian approach allows prior information to be incorporated, when appropriate; and

(6) Likelihood -based approaches like Bayes or maximum likelihood have the property of large-sample efficiency, and hence match or outperform design-based inferences if the model is correctly specified.

An alternative to a direct Bayesian modeling approach for incorporating auxiliary information is model-assisted estimation, where a model is applied to predict the non-sampled values, and then the predictions are "calibrated by applying the HT estimator to the residuals from that model (Särndal, Swensson and Wretman 1992). Specifically, the generalized regression estimator of *T* takes the form:

$$\hat{T}_{\text{gr}} = \sum_{i=1}^{N} \hat{y}_i + \sum_{i \text{ sampled}} (y_i - \hat{y}_i)/\pi_i \,, \tag{9}$$

where $\hat{y}_i$ is the prediction from a linear regression model relating Y to the covariates. While this approach is popular and yields design-consistent (Isaki and Fuller 1982) estimates, my personal preference is to choose robust models that yield design-consistent estimates, that is, to correct the model rather than to correct the estimator. It is relatively easy to find models that yield design consistent estimates (e.g. Firth and Bennett (1998), and there is little evidence that calibration yields better inferences than direct model estimates when the latter are design consistent.

A criticism of the model-based approach is that it is impractical for large-scale survey organizations: the work in developing strong models, and the computational

complexity of fitting them, is not suited to the demands of "production-oriented" survey analysis. However, attention to models is needed in model-assisted approaches, even when the basis for inference is the sample design. Also, computational power has expanded dramatically since the days of early model versus randomization debates, and much can be accomplished using software for mixed models in the major statistical packages (SAS 1992; Pinheiro and Bates 2000) or Bayesian software based on MCMC methods such as BUGS. (Spiegelhalter, Thomas, and Best 1999). Bayesian software targeted at complex survey problems would increase the utility of this approach for practitioners. Also, guidance on "off-the-shelf" models for routine application to standard sample designs would be useful, although no statistical procedure, design or model-based, should be applied blindly without any attention to diagnostics of fit to the data.

References

Box, G.E.P. (1980), "Sampling and Bayes inference in scientific modelling and robustness" (with discussion), *Journal of the Royal Statistical Society Series A* 143, 383-430.

Basu, D. (1971), "An essay on the logical foundations of survey sampling, Part 1," in *Foundations of Statistical Inferenc*e, Toronto: Holt, Rinehart and Winston, pp. 203-242.

Binder, D. A. (1982), "Non-parametric Bayesian Models for Samples from Finite Populations," *Journal of the Royal Statistical Society* 44, 3, 388-393.

Breidt, F.J. and Opsomer, J.D. (2000), "Local Polynomial Regression Estimators in Survey Sampling," *Annals of Statistics* 28, 1026-53.

Brewer, K.R.W. and Mellor, R.W. (1973), "The Effect of Sample Structure on Analytical Surveys," *Australian Journal of Statistics* 15, 145-152.

Chen, Q., Elliott, M.R. & Little, R.J. (2008). Bayesian Penalized Spline Model-Based Estimation of the Finite Population Proportion for Probability-Proportional-to-Size Samples. Submitted to *Survey Methodology*.

Cochran, W.G. (1977), *Sampling Techniques*, 3rd Edition, New York: Wiley.

Cox, D.R. and Hinkley, D.V. (1974), *Theoretical Statistics*, London: Chapman Hall.

Dumouchel, W.H. and Duncan, G.J. (1983), "Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples," *Journal of the American Statistical Association* 78, 535-543.

Elliott, M. R. and Little, R.J.A. (2000). Model- Based Alternatives to Trimming Survey Weights. *Journal of Official Statistics* 16, No. 3, 191-209.

Ericson, W.A. (1969), "Subjective Bayesian Models in Sampling Finite Populations," *Journal of the Royal Statistical Society*, B 31, 195-234.

Ericson, W. A. (1988), "Bayesian Inference in Finite Populations," in *Handbook of Statistics* 6, Amsterdam: North-Holland, pp. 213-246,

Firth, D. and Bennett, K.E. (1998), "Robust Models in Probability Sampling," *Journal of the Royal Statistical Society*, B 60, 3-21.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003), *Bayesian Data Analysis*, 2nd. edition. New York: CRC Press.

Ghosh, M. and Meeden, G. (1997), *Bayesian Methods for Finite Population Sampling*. London: Chapman & Hall.

Godambe, V.P. (1955), "A Unified Theory of Sampling from Finite Populations," *Journal of the Royal Statistical Society*, B 17, 269-278.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953), *Sampling Survey Methods and Theory*, Vols. I and II, New York: Wiley.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," *Journal of the American Statistical Association* 78, 776-793 (with discussion).

Holt, D., and Smith, T.M.F. (1979), "Poststratification," *Journal of the Royal Statistical Society*, A 142, 33-46.

Holt, D., Smith, T.M.F., and Winter, P.D. (1980), "Regression Analysis of Data from Complex Surveys," *Journal of the Royal Statistical Society*, A 143, 474-87.

Horvitz, D.G., and Thompson, D.J. (1952), "A Generalization of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association* 47, 663-685.

Isaki, C. T., and Fuller, W. A. (1982), "Survey Design Under the Regression Superpopulation Model", *Journal of the American Statistical Association* 77, 89-96.

Kish, L. (1965), *Survey Sampling*, New York: Wiley.

Kish, L. and Frankel, M.R. (1974), "Inferences from Complex Samples (with discussion), *Journal of the Royal Statistical Society* B 36, 1-37.

Konijn, H.S. ((1962), "Regression Analysis in Sample Surveys," *Journal of the American Statistical Association* 57, 590-606.

Korn, E.L. and Graubard, B.I. (1999), *Analysis of Health Surveys*, New York: Wiley.

Lazzeroni, L.C., and Little, R.J.A. (1998), "Random- Effects Models for Smoothing Post-Stratification Weights," *Journal of Official Statistics* 14, 61-78.

Little, R.J.A. (1982), "Models for nonresponse in sample surveys," *Journal of the American Statistical Association* 77, 237-250.

Little, R.J.A. (1988). Missing data in large surveys. *Journal of Business and Economic Statistics,* 6, 287-301 (with discussion).

Little, R.J.A. (1991), "Inference with Survey Weights," *Journal of Official Statistics* 7, 405-424.

Little, R.J.A. (1993), "Post-Stratification: a Modeler's Perspective," *Journal of the American Statistical Association* 88, 1001-1012.

Little, R.J.A. (2003a). The Bayesian Approach to Sample Survey Inference. In "*Analysis of Survey Data*," R.L. Chambers & C.J. Skinner, eds., pp. 49-57. Wiley: New York.

Little, R.J.A. (2003b). Bayesian Methods for Unit and Item Nonresponse. In "*Analysis of Survey Data*," R.L. Chambers & C.J. Skinner, eds., pp. 289-306. Wiley: New York.

Little, R.J.A. (2004). To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association,* 99, 546-556.

Little, R.J.A. (2006). Calibrated Bayes: A Bayes/Frequentist Roadmap. *The American Statistician*, 60, 3, 213-223.

Little, R.J.A., and Rubin, D.B. (2002), *Statistical Analysis with Missing Data,* 2nd edition, New York: Wiley.

Little, R.J. & Vartivarian, S. (2003). On weighting the rates in nonresponse weights. *Statistics in Medicine* 22, 1589-1599.

Little, R.J.A. & Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*, 31, 161-168.

Mahalanobis, P.C. (1943), "Recent Experiments in Statistical Sampling in the Indian Statistical Institute," *Journal of the Royal Statistical Society* 109, 325-378.

Pfeffermann, D. (1993), "The Role of Sampling Weights when Modeling Survey Data," *International Statistical Review* 61, 317-337.

Pfeffermann, D. and Holmes, D.J. (1985), "Robustness Considerations in the Choice of Method of Inference for Regression Analysis of Survey Data," *Journal of the Royal Statistical Societ*y, A 148, 268-278.

Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-Plu*s, New York: Springer.

Rao, J.N.K (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

Rao, J.N.K. (1997). Developments in sample survey theory: an appraisal. *Can. J. Statist.* 25, 1-21.

Rao, J.N.K. (2003). *Small Area Estimation,* New York: Wiley, 313 pp.

Royall, R. M. (1970), "On Finite Population Sampling Under Certain Linear Regression Models," *Biometrika* 57, 377-387.

Rubin, D.B. (1976), "Inference and Missing Data," *Biometrika* 53, 581-592.

Rubin, D.B. (1978), Bayesian Inference for Causal Effects: The Role of Randomization. *Annals of Statistics* 6(1), 34-58.

Rubin, D.B. (1984), "Bayesianly justifiable and relevant frequency calculations for the applied statistician", *Annals of Statistics* 12, 1151-1172.

Rubin, D.B. (1983), Comment on "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," by M.H. Hansen, W.G. Madow, and B.J. Tepping, *Journal of the American Statistical Association* 78, 803-805.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Survey*s. Wiley: New York.

Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992), *Model Assisted Survey Sampling*, Springer Verlag: New York.

SAS (1992), "The Mixed Procedure," in *SAS/STAT Software: Changes and Enhancements, Release 6.07,* Technical Report P-229, SAS Institute, Inc., Cary, NC.

Scott, A.J. (1977), "Large-Sample Posterior Distributions for Finite Populations," *Annals of Mathematical Statistic*s, 42, 1113-1117.

Scott, A.J. and Smith, T.M.F. (1969), "Estimation in Multistage Samples," *Journal of the American Statistical Association* 64, 830-840.

Smith, T.M.F. (1988), "To Weight or not to Weight, that is the Question," in *Bayesian Statistics* 3, J.M. Bernado, M.H. DeGroot and D.V. Lindley, eds., Oxford, U.K.: Oxford University Press, pp. 437-451.

Spiegelhalter, D.J., Thomas, A. and Best, N.J. (1999), *WinBUGS Version 1.2 User Manua*l, MRC Biostatistics Unit, Cambridge, UK.

Thompson, M.E. (1988), ``Superpopulation Models", *Encyclopedia of Statistical Sciences (Vol. 1)*, 9**,** 93-99 .

Valliant, R. , Dorfman, A.H. , and Royall, R. M. (2000), *Finite Population Sampling and Inference: a Prediction Approac*h, New York: Wiley.

Zheng, H. & Little, R.J. (2003). Penalized Spline Model-Based Estimation of the Finite Population Total from Probability-Proportional-To-Size Samples. *Journal of Official Statistics*, 19, 2, 99-117.

Zheng, H. & Little, R.J. (2004). Penalized Spline Nonparametric Mixed Models for Inference about a Finite Population Mean from Two-Stage Samples. *Survey Methodology*, 30, 2, 209-218.

Zheng, H. & Little, R.J. (2005). Inference for the Population Total from Probability-Proportional-to-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model. *Journal of Official Statistics*, 21, 1-20.